Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 4918–4922, June 1992
Biophysics

Optimal protein-folding codes from spin-glass theory

(protein-structure prediction/associative memories/neural networks/biomolecular dynamics)

RICHARD A. GOLDSTEIN*, ZAIDA A. LUTHEY-SCHULTEN*†, AND PETER G. WOLYNES*‡

*School of Chemical Sciences, †National Center for Supercomputing Applications, and ‡Beckman Institute, University of Illinois, Urbana, IL 61801

Contributed by Peter G. Wolynes, February 14, 1992

ABSTRACT **Protein-folding codes embodied in sequence-dependent energy functions can be optimized using spin-glass theory. Optimal folding codes for associative-memory Hamiltonians based on aligned sequences are deduced. A screening method based on these codes correctly recognizes protein structures in the "twilight zone" of sequence identity in the overwhelming majority of cases. Simulated annealing for the optimally encoded Hamiltonian generally leads to qualitatively correct structures.**

The problem of protein-structure prediction from sequence has been described as the determination of the second half of the genetic code (1). Unlike the code translating DNA sequences into amino acid sequences, the mathematical structure of the folding code remains problematic because of the differing character of sequential and structural information. Although it is conceivable that a strictly deterministic local code exists, there is considerable evidence that the folding code is fuzzy and nonlocal. A particularly simple representation of a class of codes is provided by the associative-memory Hamiltonians introduced by Friedrichs and Wolynes (2). The associative-memory Hamiltonian encodes correlations between the sequence of the target protein whose structure is to be determined and a set of memory proteins $\{\mu\}$ as well as the structures of the memory proteins through sets of pair distances. Minimization of the associative-memory Hamiltonian yields the predicted structure of the target protein. The associative-memory Hamiltonians resemble empirical energy functions long used for proteins, but their form was motivated by the theory of neural networks (3). This allows the use of the ideas developed for pattern recognition by neural networks and thermodynamic formulations of information processing based on spin-glass theory. Many aspects of the relevance of spin-glass theory to folding phenomenology have already been explored (4–9). In this paper we show how spin-glass theory can be used to optimize associative-memory Hamiltonians and lead to a characterization of optimal protein-folding codes.

When sequence homology of a new protein with a protein of known structure is high, standard alignment techniques allow the prediction of structure. The exact degree of homology necessary to make the inference depends upon standards of homology and structural similarity but is in the range of 25–40% sequence identity (10, 11). Below this limit, a "twilight zone" emerges in which sequence homology does not imply structural similarity (11). Our method of searching for optimal protein-folding codes focuses on the twilight zone of sequence identity.

We use a random-energy approximation to the thermodynamics of associative-memory Hamiltonians (4, 5). Finding which of the configurations of the memory-protein structures is the most stable leads to a screening method for recognizing protein structures. In contrast to other approaches (12–14), the screening method directly discriminates between many structures in determining the assignment. Optimal protein-folding codes, deduced for proteins characterized by structural class, are very successful in such a screening procedure.

A stringent test of the energy function is provided by molecular dynamics. Energy functions based on optimal codes usually lead to minima recognizably similar to the correct protein structures. In some cases the similarity is extraordinarily good, while in others there is overcollapse in sections of the protein leading to structures with reasonably good distance matrices but with stereochemical irregularities. These results are described below.

Code Optimization Using the Random-Energy Approximation

In its simplest form, the associative-memory (AM) Hamiltonian as a function of the pairwise distance between the α-carbons of residues i and j, r_{ij}, has the form

$$\mathcal{H}_{AM} = -\sum_{\mu} \sum_{i<j} \gamma_{ij}^{\mu} \theta(r_{ij} - r_{ij}^{\mu}) + \mathcal{H}_o. \qquad [1]$$

γ_{ij}^{μ} encodes a degree of similarity between residues i and j of the target protein and memory protein μ. $\theta(r_{ij} - r_{ij}^{\mu})$ is a Gaussian function of the difference between the pairwise distance in the target structure and the memory structure, and $\mathcal{H}_o$ is a typical chain molecule Hamiltonian for the backbone atoms. Various forms of backbone Hamiltonian have been investigated (15). The form of γ_{ij}^{μ} may include information about the probability of mutation of the various residues in the pair ij, their physicochemical properties, or the context of the residues in the protein as represented by predicted secondary structure (15). As there may be insertions and deletions in the target sequence compared with the memory protein, the sequence number i and j of corresponding residues may differ. In this paper, we use standard alignment techniques to match corresponding residues, although other methods of generalization have been used (15). The alignment indicates which residues, designated i' and j', are related to i and j in the target protein. For aligned sequences, γ_{ij}^{μ} in Eq. 1 is replaced by $\gamma_{ii'j'}^{\mu}$ and r_{ij}^{μ} is replaced by $r_{i'j'}^{\mu}$. Prealigning the sequences induces correlations between the sequences and structures of the target protein and the unrelated memories, invalidating the intuition behind our prior choices for folding codes and necessitating folding-code optimization.

The phase diagram of the generic associative-memory Hamiltonian has been studied by Sasai and Wolynes (16). This reveals the underlying competition between two different phase transitions. If the data base of structures used to develop the code is small and the association between the sequence of the protein to be folded and an example is high, the minimization of the associative-memory Hamiltonian by molecular dynamics will lead to a first-order phase transition at a folding temperature, T_f. On the other hand if the number

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: Brookhaven PDB, Brookhaven Protein Data Bank.

of different possible structures is large and the relationship of sequences is small, the energy landscape is rough, and the system undergoes a thermodynamic glass transition at a temperature T_g. Close to T_g the protein dynamics is slow and simulated annealing fails. A reasonable figure of merit for a particular γ is the ratio of the folding temperature to the glass transition temperature, T_f/T_g. This ratio should be large so that annealing can be carried out at high temperature, where the trapping in local minima will not be important. Convenient approximations to T_f/T_g allow the rapid search through possible choices of γ. Maximizing T_f/T_g is a practical implementation of the principle of minimal frustration (4).

A simple analysis of the competition between folding and the glass transition can be made by using the random-energy approximation (4, 5). In this approximation, the energy levels of the liquid-like states of the protein are defined as random and correlations between states are neglected. According to Bryngelson and Wolynes, the distribution of energy levels will depend on the fraction of native structure. Their analysis can be simplified if we assume that the liquid-like state has virtually no native structure, while the folded state is nearly perfect. The entropy of the folded state may then be neglected. However, the liquid-like phase has high entropy S_o and a wide standard deviation of energies, δE. To a first approximation, the folding temperature is given by

$$T_f = \frac{\Delta E + (\Delta E^2 - 2S_o\delta E^2)^{1/2}}{2S_o}, \quad [2]$$

where ΔE is the average energy difference between the folded and liquid-like states.

The glass transition occurs when the entropy difference between the folded and unfolded states vanishes. In this model, the transition temperature, T_g, is given by $T_g = (\delta E^2/2S_o)^{1/2}$. One can then show that T_f/T_g is maximized when $\lambda = \Delta E/\delta E$ is maximized.

If the structures in the liquid state are minima for $\mathcal{H}_o$, ΔE is linear in γ while δE^2 is quadratic in γ, or in matrix form $\Delta E = \mathbf{A}\gamma$ and $\delta E^2 = \gamma\mathbf{B}\gamma$. Apparently the maximization of λ presents a nonlinear problem with possible multiple minima, but since γ is arbitrary up to multiplication by a constant, the maximization of T_f/T_g leads to the explicit form for $\gamma{:}\gamma = \mathbf{B}^{-1}\mathbf{A}$.

A and **B** may be evaluated for any choice of functional form for γ, given correct structures and liquid-like model structures. Various forms for γ can be compared based on resulting values of λ.

Results

Optimal Codes Based on Hydrophobicity. The result of the random-energy model for γ is reminiscent of the pseudo-inverse approximations in feed-forward neural nets (17). Just as there, simple forms of γ are the best for generalization. In this paper we discuss results obtained using the association constants dependent only on the hydrophobicity of the residues in the target and example proteins and on the distances between the residues. We use a binary representation of the Eisenberg consensus hydrophobicity scale to label the sequence (18) as described earlier (2, 15, 19).

In the first code, γ is independent of distance, resulting in a Hamiltonian dominated by long-range interactions. In addition, we construct a γ vector in which independent γ values are assigned to residues close in sequence ($j - i < 5$) and close in space ($r_{ij} < 8, j - i > 5$).

We used the x-ray coordinates of a set of proteins between 10% and 50% larger than the target protein, including all possible translations along the sequence, to model the liquid-like states. The average and variance of the energy of the target protein in these configurations is used to calculate **A** and **B**.

A set of 123 target proteins 50–250 residues long, chosen from the Brookhaven Protein Data Bank (PDB), was prepared (20, 21). These target proteins were aligned to a data set of 185 proteins and protein subunits having well-defined structures. The alignments were obtained by using the Best-Fit algorithm of the Genetics Computer Group (GCG; Madison, WI) package with default specifications (22). For each

Table 1. Predictions with the screening method

Protein	PDB	% I	q
	α-helix		
Human fetal hemoglobin	1FDH(G)	27.90	0.70
Sickle-cell hemoglobin	1HDS(A)	32.80	0.64
Sickle-cell hemoglobin	1HDS(B)	28.40	0.78
Sea-hare myoglobin	1MBA	32.60	0.53
Phage 434 repressor	1R69	28.60	0.67
Phage 434 Cro protein	2CRO	16.40	0.46
Human hemoglobin	2HHB(A)	35.90	0.78
Human hemoglobin	2HHB(B)	26.50	0.70
Lupine leghemoglobin	2LH4	23.60	0.52
Sea lamprey hemoglobin	2LHB	35.90	0.67
P. aeruginosa cytochrome 551	351C	28.00	0.51
R. rubrum cytochrome c_2	3C2C	40.00	0.67
Bovine Ca-binding protein	3ICB	32.30	0.49
Tuna cytochrome *c*	5CYT	40.00	0.79
Sperm whale myoglobin	5MBN	27.90	0.70
	β-sheet		
Mouse R19.9 Fab fragment	1F19(H)	29.60	0.44
Mouse R19.9 Fab fragment	1F19(L)	26.20	0.52
Human *Rhinovirus* coat	1R1A(3)	34.20	0.70
Human Bence–Jones fragment	1REI(A)	30.20	0.69
Human Fab fragment	2FB4(H)	31.40	0.51
Human Fab fragment	2FB4(L)	31.40	0.57
Mouse IgA Fab fragment	2FBJ(L)	26.20	0.52
Monkey mengovirus coat	2MEV(2)	37.10	0.63
Monkey mengovirus coat	2MEV(3)	34.20	0.73
Human *Poliovirus* coat	2PLV(3)	32.40	0.60
Human Bence–Jones fragment	2RHE	30.90	0.59
Mouse IgG1 Fab fragment	3HFM(H)	26.20	0.42
Mouse IgG1 Fab fragment	3HFM(L)	27.70	0.43
Human histocompatibility Ag	3HLA(B)	26.50	0.65
Mouse IgG2 Fab fragment	4FAB(H)	28.10	0.51
Mouse IgG2 Fab fragment	4FAB(L)	25.70	0.43
Human *Rhinovirus* 14 coat	4RHV(3)	27.00	0.70
Poplar plastocyanin	5PCY	25.80	0.55
	Mixed category		
Baboon α-lactalbunin	1ALC	38.70	0.80
D. vulagis flavodoxin	1FX1	31.10	0.62
Human lysozyme	1LZ1	38.70	0.70
Human DHFR	2DHF(A)	31.20	0.55
Hen lysozyme	2LYZ	37.20	0.74
B. subtilis eglin C	2SEC(1)	35.50	0.82
Barley chymotrypsin inhibitor	2SNI(I)	37.10	0.74
L. casei DHFR	3DFR	30.60	0.74
Cl. MP flavodoxin	3FXN	31.10	0.70
E. coli DHFR	4DFR(B)	34.00	0.72
Chicken DHFR	8DFR	34.00	0.45

Results of the screening method with structuraly dependent hydrophobicity/proximity codes. The Brookhaven PDB designation with substrand identity in parentheses is listed in the second column. The third column lists the percentage identity (% I) between aligned sequences of the target and the most homologous structure partner. The next column gives the structural similarity of the predicted structure compared with the correct structure, as measured by q values (Eq. 3). Organisms from top to bottom are: *Pseudomonas aeruginosa, Rhodospirillum rubrum, Desulfovibrio vulgaris, Bacillus subtilis, Lactobacterium casei, Clostridium MP*, and *Escherichia coli*. DHFR, dehydrofolate reductase.

target protein, the 25 most homologous proteins were chosen as memory proteins, excluding proteins having more than 40% sequence identity with the target protein. For our data base, this cutoff represented the start of the region where sequence homology did not necessarily mean structural similarity.

Structural similarities were based on q scores.

$$q^{T\mu} = [N(N-1)]^{-1} \sum_{i \neq j} \theta(r_{ij} - r^{\mu}_{i'j'}) \qquad [3]$$

is the q score for target T and memory μ. q scores in excess of 0.4 indicate structural similarity. This cutoff corresponds to rms deviations generally less than 6 Å. With this criterion, 31 of the target proteins had at least one memory that was structurally similar. These constituted the initial target set.

One seeks a universal γ that incorporates correlations present over the entire data set yet is independent of the coordinates of the target structure. We employed a jack-knife method, averaging **A** and **B** over the set of target proteins excluding the one to be predicted.

The encodings were obtained for the purpose of carrying out determinations of structures using molecular dynamics, but a simpler screening method can be used. If the correctly folded state is close to the configuration of at least one of the memory proteins, we can calculate the energy of the protein in the configuration corresponding to each of the memory proteins in turn. The lowest-energy structure represents the prediction.

The global code based on hydrophobicity performs excellently for the 17 β-sheet and mixed-category proteins, choosing a correct structure 94% of the time for both functional forms of γ, failing only for plastocyanin (Brookhaven PDB code 5PCY). The α-helical performance was poorer, with a structurally similar memory protein chosen for only 64% or 71% of the target proteins, with the γ based on simple hydrophobicity or hydrophobicity/proximity, respectively. λ values tend to be lower in this class. Even with the α-helical proteins, this screening method was superior to the assumption that the most homologous protein was structurally similar, which is true for only 43% of the α-helical targets.

Table 2. Predictions of twilight zone proteins

PDB	% I	q
	α-helix	
1FDH(G)	22.90	0.56
1HDS(A)	18.80	0.61
1HDS(B)	24.80	0.78
1MBA	23.60	0.53
2CRO	16.40	0.46
2HHB(A)	21.80	0.50
2HHB(B)	25.00	0.81
2LH4	23.60	0.53
2LHB	18.80	0.13
351C	22.20	0.30
5MBN	25.00	0.70
	β-sheet	
1F19(H)	24.70	0.44
1F19(L)	24.90	0.44
2FB4(H)	25.00	0.46
2FBJ(L)	25.00	0.50
3HFM(H)	23.60	0.42
3HFM(L)	23.90	0.30
3HLA(B)	24.70	0.65
4FAB(H)	24.90	0.54
4FAB(L)	22.50	0.41

Results of the screening method to predict structures for proteins of low (<25%) homology. The column labels are as in Table 1.

There exist algorithms to assign proteins to structural class. Because of the distinction seen between the α-helical class and the others, we optimized independently for each class, using only memory proteins corresponding to the class of the target. Excluding examples of different structural classes increases the number of target proteins with a structural similarity in the memory set to 44. The division into classes considerably improves the performance. With distance-independent encodings, the success rate is 80% in the α-helical class, 94% in the β-sheet class, and 100% in the mixed-category class. As can be seen in Table 1, when the γ code incorporating hydrophobicity and proximity is applied, one obtains a 100%-correct assignment of the protein to a structural homolog in all categories. The β-sheet proteins have especially favorable values of T_f/T_g. Assignment can be done even when the % sequence identity is as low as 16.4%, as witnessed by the Cro protein (2CRO).

Most of the targets in this set had a sequence identity of 30–40% with a structurally similar memory protein. To see how much homology is actually required for this screening method, we eliminated from the memory set all structurally similar proteins with more than 25% homology with the target. Twenty proteins still had structurally similar proteins in their respective structural classes. The γ values were optimized for these target sets, again using the jack-knife method. As shown in Table 2, the screening method yields a structurally similar protein for 9 of the 11 α-helical proteins and 8 of the 9 β-sheet ones. For all of the failures except 2LHB, the prediction was still of the same family as the target.

Simulated Annealing. Nine proteins with both moderate to low T_f/T_g, based on structural-class-dependent codes encoding hydrophobicity and proximity, were annealed by molecular dynamics. A simple harmonic backbone containing α and

Table 3. Results of simulated annealing

	ΔE	Best memory			Predicted structure	
PDB	δE	% I	q	rms	q	rms
		α-helix				
2CRO	3.82	16.4	0.46	4.21	0.32	6.06
					0.32	6.35
351C	4.98	28.0	0.51	5.43	0.28	5.41
3ICB	10.89	32.3	0.49	3.15	0.40	5.13
5CYT	8.99	40.0	0.79	2.04	0.23	9.95
					0.24	10.14
		β-sheet				
1REI(A)	20.42	30.2	0.69	2.82	0.61	2.68
					0.59	3.10
3HLA(B)	17.27	26.5	0.66	2.89	0.59	5.15
5PCY	11.86	25.8	0.55	3.78	0.27	8.61
					0.31	7.34
		Other				
1ALC	13.98	38.7	0.83	1.98	0.47	7.58
					0.47	7.81
2SNI(I)	14.06	37.1	0.80	1.64	0.40	5.54
					0.41	6.43
					0.34	8.72

Results of simulated annealing for nine targets. The Hamiltonian was based on a γ^{μ}_{ij} encoding both hydrophobicity and proximity optimized for the structural class of the target. The PDB designation of the protein is listed in column 1. Column 2 lists the value of $\lambda = \Delta E/\delta E$ for the target. The next three columns list the degree of sequence identity of the target with the most homologous structure partner and the q score and rms deviation of the most similar example. The last two columns give q scores and rms deviation with respect to the correct structure. For multiple runs, the results are listed in increasing order of final energy. Column labels are as in Table 1.

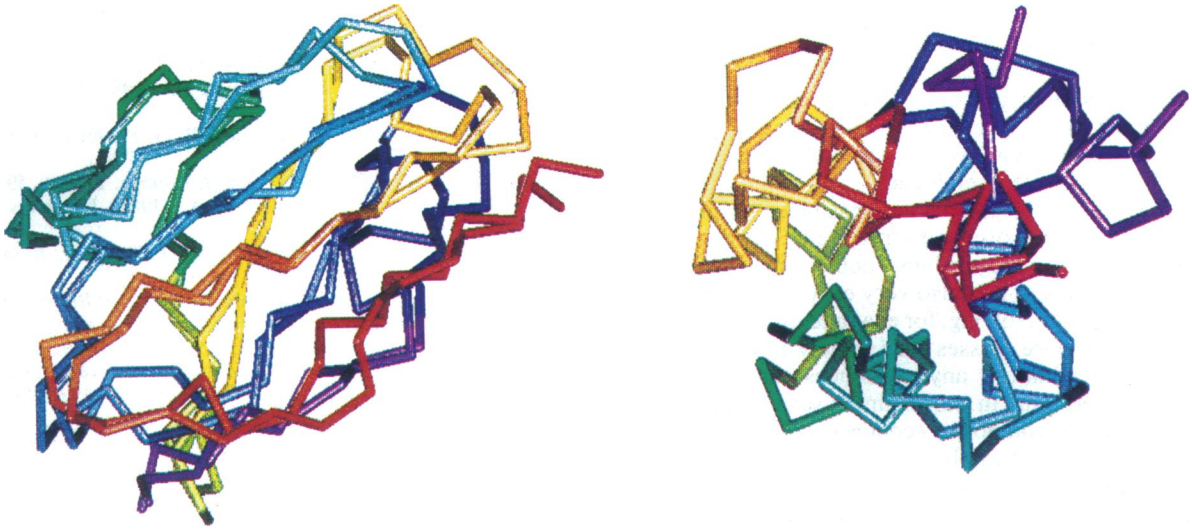

FIG. 1. Predicted structures of 1REI (*Left*) and 2CRO (*Right*) compared with their respective native structures. Proteins are colored identically by sequence number, with the first residue colored red and the last purple.

β carbons was used, with identical γ for all types of interactions. The annealing protocols described in our earlier paper (19) were used. Each run takes approximately 1 hr on a CRAY 2 supercomputer. As shown in Table 3, the prediction was topologically similar to the correct structure for 8 of the 9 targets. Proteins with λ values greater than about 12 give structures of q greater than 0.4. An example from this range is the variable fragment from a Bence–Jones immunoglobulin [1REI(A)] with a λ value of 20.4 (Fig. 1). The runs with lower λ values still give recognizably correct structures in most cases. Cro protein from phage 434 has only 16% sequence identity and a λ value of 3.82, yet the prediction and target structure compare well (Fig. 1).

Some distance matrices for the actual and predicted structures are shown in Fig. 2. The main elements of the distance matrix are often quite close, but the prediction has a tendency to be overcollapsed in places. The only structure that failed in this set is tuna cytochrome *c* (5CYT). The predicted structure has a q of only 0.24. Even here, much of the topology is preserved (see distance map). Possibly the pattern of amino acid residues around the heme group is not well represented by the binary hydrophobicity scale.

The errors in prediction often involve violations of stereochemical rules or excluded volume constraints. More detailed backbone models may improve performance.

Conclusion

The results in this paper show that preprocessing of sequences by standard alignment methods and code optimization using spin-glass theory considerably improves performance in predicting protein structure. Two different algorithms for tertiary structure recognition from sequences with identity in the twilight zone were shown. The screening

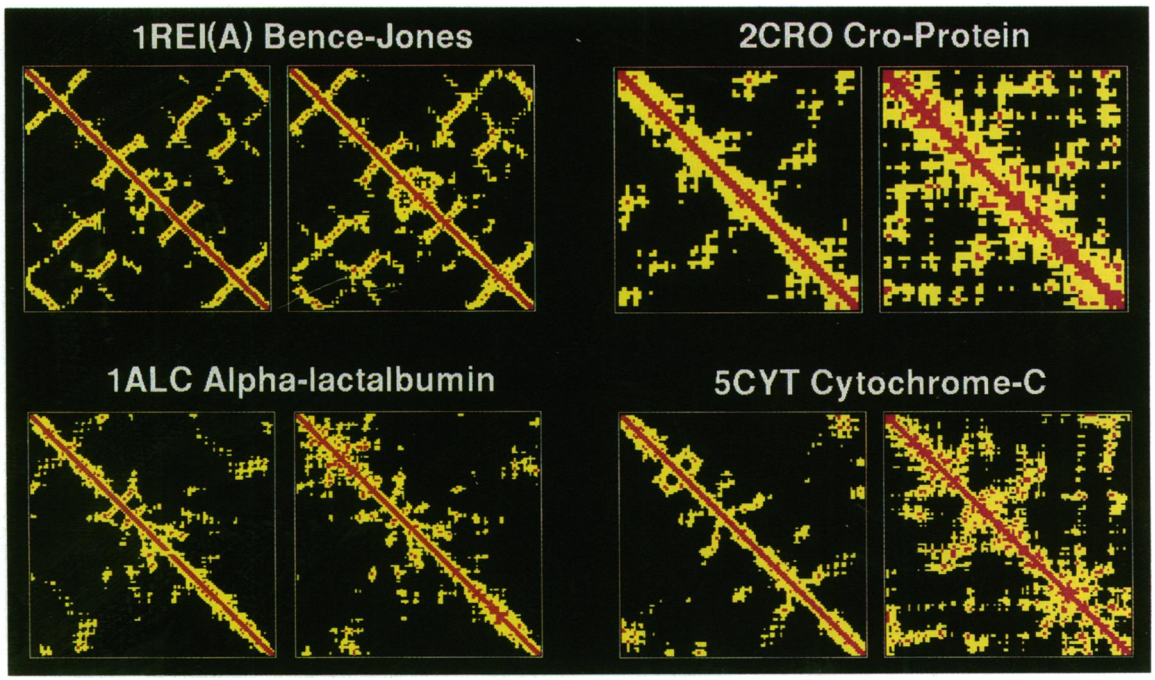

FIG. 2. Distance maps for the crystal structure (*Left*) and the simulated annealing result (*Right*) for four target proteins. The residue number index increases from top to bottom and from left to right. A point for a pair of residues whose α carbons are within 10 Å is yellow and within 5 Å is red. The distance maps show similar features except for 5CYT (PDB code).

method gives a highly reliable scheme for discriminating between possible putative structures, especially if the structural class of the protein can be predetermined. Structure determination by simulated annealing is also generally successful. A most dramatic example is the determination of a structure of Cro protein, which has only 16% sequence identity with its closest homologue.

Several further developments are possible. The alignment method we used was chosen conservatively. The continuing advances in alignment algorithms should allow even more efficient prescreenings. It is also very easy to evaluate more elaborate encodings involving, for example, finer divisions of amino acids into more classes than merely hydrophobic or hydrophylic. In addition, any sequences that are highly homologous to proteins in our example set can be used to build consensus or composite sequences. The appropriate weighting of different examples can be approached using the same optimization techniques.

The optimization approach may also be used for Hamiltonians based on feature detection, rather than using prealigned sequences. Such Hamiltonians more closely resemble physical reality. The methodology of the random energy approximation may also be used to refine Hamiltonians outside the associative memory framework. Continued development of these techniques will allow even greater accuracy in structure prediction and give insight into the dominant interactions in folding.

Helpful interactions with H. Bohr, B. Ramirez, and K. Webb and critique of the text by Hans Frauenfelder and J. A. McCammon are gratefully acknowledged. Computations were carried out at the National Center for Supercomputing Applications in Urbana. This work was supported by National Institutes of Health Grant PHSIRO/GM44557-01.

1. Gierasch, L. M. & King, J., eds. (1990) *Protein Folding: Deciphering the Second-Half of the Genetic Code* (Am. Assoc. for the Advancement of Science, Washington).
2. Friedrichs, M. S. & Wolynes, P. G. (1989) *Science* **246,** 371–373.
3. Hopfield, J. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2554–2558.
4. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
5. Bryngelson, J. D. & Wolynes, P. G. (1990) *Biopolymers* **30,** 171–188.
6. Garel, T. & Orland, H. (1988) *Europhys. Lett.* **6,** 597–601.
7. Shakhnovich, E. I. & Gutin, A. (1988) *Europhys. Lett.* **8,** 327–332.
8. Shakhnovich, E. I. & Gutin, A. (1989) *Stud. Biophys.* **132,** 47–56.
9. Wolynes, P. G. (1991) in *Spin Glasses and Biology,* ed. Stein, D. (World Scientific Press, New York), in press.
10. Sanders, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
11. Doolittle, R. F. (1990) *Methods Enzymol.* **183,** 99–110.
12. Bowie, J. U., Luthy, R. & Eisenberg, D. (1991) *Science* **253,** 164–170.
13. Luthy, R., McLachlan, A. D. & Eisenberg, D. (1991) *Proteins* **10,** 229–239.
14. Finkelstein, A. & Reva, B. (1991) *Nature (London)* **351,** 497–499.
15. Friedrichs, M. S., Goldstein, R. A. & Wolynes, P. G. (1991) *J. Mol. Biol.* **222,** 1013–1034.
16. Sasai, M. & Wolynes, P. G. (1990) *Phys. Rev. Lett.* **65,** 2740–2743.
17. Ritter, H., Martinetz, T. & Schulten, K. (1991) *Neural Computation and Self-Organizing Maps: An Introduction* (Addison-Wesley, New York).
18. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. (1982) *Faraday Symp. Chem. Soc.* **17,** 109–120.
19. Friedrichs, M. S. & Wolynes, P. G. (1991) *Tetrahedron Comput. Methodol.* **3,** 175–190.
20. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
21. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications,* eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Commission of the International Union of Crystallography, Bonn), pp. 107–132.
22. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.